Explicit reasons, implicit stereotypes and the effortful control of the mind

Explicit Reasons, Implicit Stereotypes and the Effortful Control of the Mind

Tillmann Vierkant & Rosa Hardt

School of Philosophy Psychology and Language Sciences

University of Edinburgh

Mail: t.vierkant@ed.ac.uk

Keywords: Explicit reasons, implicit biases, moral responsibility, assertibility, control, mental actions

Abstract:

Research in psychology clearly shows that implicit biases contribute significantly to our behaviour. What is less clear, however, is whether we are responsible for our implicit biases in the same way that we are responsible for our explicit beliefs. Neil Levy has argued recently that explicit beliefs are special with regard to the responsibility we have for them, because they unify the agent. In this paper we point out multiple ways in which implicit biases also unify the agent. We then examine Levy's claim that the assertibility of explicit beliefs means that they have a unique way of unifying the agent by being available for syntactical operations. We accept that syntactical operations are important, but worry that they are less straightforwardly connected to the unification of agents than Levy claims.

Intuitively it seems very plausible that our explicitly held attitudes are in a special sense relevant for moral responsibility – in a way that other attitudes, such as gut feelings and implicit biases, are not. However, this view is challenged by findings from research in the cognitive sciences, which seem to suggest that we are systematically mistaken about the attitudes that we really hold. Matt King & Peter Carruthers (2012) have argued that these findings demonstrate that conscious explicit attitudes cannot be what makes us responsible agents, because there are no conscious explicit attitudes that have the right causal profile to control our behaviour (e.g. decisions, intentions).

In reply to this work, Neil Levy (2012) has tried to articulate in which sense consciously held attitudes could be special for moral responsibility after all. It is important to note that for Levy, it is not consciousness that separates explicit reasoning from implicit attitudes, since we may be aware of our implicit biases as 'gut feelings'. Levy suggests that a better way to understand the difference is that while implicit processes and attitudes may be propositionally *described*, only explicit beliefs are propositionally *expressed*. For example: We can *describe* our implicit biases as being processes that cause us to act *as if* we were sexist, while we *express* the explicit belief that we think men and women are equals. Importantly, in this respect gut feelings (like e.g. phobias) and implicit biases fall on the same side of the explicit / implicit divide. We are very much aware of our spider phobia for example, but this does not mean that the phobia expresses our belief that spiders are dangerous, just as our knowledge that we have an implicit bias against some group does not express a negative belief about that group. Obviously, there are many important differences between the two categories. We are e.g. aware of gut feelings introspectively, whereas we know about our biases often only because of a theoretical approach to our own psychology, but for our purposes it is crucial that they both do not have what according to Levy is critical for our explicit beliefs i.e. they do not express beliefs and are not integrated into a web of reasons. [1]

Levy argues that consciously held attitudes that are explicit in this sense unite us as moral agents in a special way that other attitudes do not. Levy's strategy to argue in favour of this claim is two-pronged. On the one hand, he analyses the structural differences between explicit and implicit states and argues that implicit states cannot contribute to the agent in the same way that explicit states can because of a number of structural differences between them. On the other hand, Levy discusses a very specific mechanism, i.e. the syntactic manipulation of asserted beliefs, which only exist for explicit content.

We think that Levy's implicit / explicit distinction marks an intuitively important boundary and we will accept it as an assumption for the purposes of our argument. Starting from this assumption, we then demonstrate that neither of Levy's two strategies to show that explicit attitudes so understood have a special status for moral responsibility is successful. In part 1 we tackle Levy's structural claims. While this paper follows the standard view of moral responsibility that for a cognitive process to contribute to moral agency it must be reason-responsive and contribute to the unity of

[1] We discuss the importance of this distinction in part two.

the agent, we disagree with Levy that only explicit attitudes meet these criteria. We argue that because implicit biases are also reason-responsive and contribute to an agent's unity, we should be held morally responsible for them as well.

In part 2, we explain Levy's focus on the syntactical operations available to explicit belief formation in terms of an underlying belief that these operations give us intentional control, something it seems we lack when it comes to implicit biases. We argue that there is no unproblematic way for Levy to characterise the intentional control of novel belief formation as the distinguishing feature of explicit beliefs, while also making it clear how implicit biases and explicit beliefs differ in how they contribute to moral responsibility.

PART 1 Implicit biases and unified agents

The Uniqueness of Explicit Reasons

Levy (2012) argues that only explicit mental processes are capable of contributing to us as unified agents. This is important, because Levy argues that moral responsibility depends on the diachronicity of the agent. So this, in turn, is why we are held responsible for mental states that unify us. Levy writes:

"Our mental states [must] be relatively unified, so that we are able to pursue our plans across time without later person-stages undercutting the projects of earlier person-stages. The kind of unity imposed on a life by implicit attitudes does not enable a person to pursue projects across time." (Pp.19).

We agree that being a unified agent is a condition for being held responsible for one's attitudes and actions, and we also agree that mental processes that allow us to act consistently through time do have a special relevance for moral responsibility. However, unlike Levy, we think that implicit biases and gut feelings can also contribute to us as coherent diachronic agents, and therefore belong to the category of states that we are responsible for.

Having clarified Levy's claim and our worry about it, we can now look at his argument about why implicit states cannot fulfill this role in more detail and explain why we disagree.

What are implicit biases?

Implicit biases are automatic stereotypes and attitudes concerning members of a social category that appear dissociable from our explicit beliefs. Implicit biases are currently understood as referring to the strength of an automatic association between either a social category and particular semantic content, known as an implicit stereotype, or between a social category and an evaluation of like / dislike or favourable / unfavourable, referred to as an implicit attitude (Greenwald & Krieger, 2006). We have an implicit stereotype of homosexual men if we automatically associate them with 'style' more than heterosexual men. We have a negative implicit attitude towards homosexual men if we automatically evaluate them less favourably than heterosexual men. Implicit Association Tests (IATs) test these associations, and their results systematically predict our behaviour. For example, implicit biases have been shown to

predict how positive one's body language is towards people of a particular social category (Dovidio et al. 2002).

Levy rejects implicit biases as part of our morally evaluable selves through noting several interdependent binaries that separate our implicit and explicit cognitive mechanisms. Levy thinks that only explicit beliefs can be rational, in the sense that they can be subjected to syntactical procedures such as those of formal logic. According to him, this means that our explicit beliefs have greater coherence than our implicit processes, as syntactical procedures give us the capacity to prevent, detect and remove contradictions.

Levy also thinks that explicit reasons are structured differently from implicit beliefs. While explicit processes have a broad content with a tight structure, implicit processes have a narrow content with a loose structure. By 'broad' content, Levy means that explicit beliefs have content that responds to a variety of features and situations. It is unclear what Levy means by a 'tight' structure. However, one way to understand 'tight' is that the content is formed of integrated and systematically related parts.

Finally, Levy suggests that only explicit attitudes normatively constrain our dispositions. That is, they lead us to "pursue behaviours which are consistent with that attitude" (p. 7). To be 'normatively constrained' by our attitudes suggests that they prescribe how we *should* behave.

The output of these features is a unified agent. Because explicit attitudes have a broad, tight, and coherent structure, and they normatively constrain our dispositions, they produce integrated, coherent and broad behaviours: an agent that can pursue projects through time. We believe, however, that implicit biases perform the same function, and explain how in the rest of part 1.

The Structure of Implicit Biases

Levy posits implicit biases as having an incoherent structure, but we will argue that they nevertheless have the right kind of structure to unify the agent. Later, we will show that this structure constrains our behaviour in a rationally consistent way.

To start with, implicit biases *do* appear to have broad content, in the sense that they respond to a variety of features and situations. For example, implicit racial biases have been demonstrated to affect our behaviour in various situations: How U.S. citizens voted in the 2008 elections (Greenwald et al., 2009a), how likely we are to trust someone of a different race (Stanley et al., 2011) and the extent to which we react with empathy when we see certain racial groups in pain (Avenanti, Sirigu & Aglioti, 2010).

Levy concedes that implicit biases are a type of implicit process that has broader content than other implicit attitudes (e.g. phobias), but argues that, since they are associative, their structure still lacks the potential to unify us.

What is wrong with the structure of our implicit biases? Levy thinks they are 'loosely' structured, and not 'tightly' structured like our explicit attitudes. At points, Levy links this to the associative, rather than rule-based, structure of implicit biases. Although

we agree that implicit biases can be understood as associatively structured, it is not clear why an associative structure causes an attitude to be 'loose' and why that would matter. Levy could believe that they lack integration, systematicity and consistency. This structure seems to be what is needed for an individual to be unified diachronically, such that, at any particular point in time, they don't contradict themselves.

One may think that what Levy means by implicit biases being 'loose' is not that their associative structures lack systematicity and consistency. Perhaps he just means that explicit reasoning is open to inferential processing in a way that implicit biases are not. We are not disputing this claim. However, as Gigerenzer (2008) points out, our explicit views are much less frequently subjected to rule-based inferential processing than we imagine and even if inferential processing is used, it is often far less successful than relying on gut feelings.

On the other hand, it seems clear that implicit biases can consist of a number of systematically integrated parts. For example, our implicit gender stereotypes associate women with being communal, egalitarian, passive, submissive, domestic, and artistic whereas men are understood as agentic, hierarchical, dominant, assertive, mathematical, scientific and more likely to be leaders (Nosek, Banaji & Greenwald, 2002; Project Implicit online, 2011; Rudman and Glick, 2001; Schmid Mast, 2004). This content is no less complex than our explicit stereotypes, when we have them.

Furthermore, the different features within implicit biases are systematically related as they often relate to the social and economic roles of members of a social category. For example, the linking of passivity, communality and domesticity with women forms a coherent and integrated whole, because these words are systematically linked to their place in society. This is supported by a study by Hoffman and Hurst (1990). They created a fictional society composed of 'Orinthians', child raisers, and 'Ackmians', city workers. Participants judged the Orinthians to be kind, patient and understanding, while the Ackmians were forceful and self-confident. These descriptions chime with the implicit stereotypes of women as 'communal' and men as 'agentic'. It is not just role, but also the amount of status a group has which affects the stereotype of them. There is a noted pattern of stereotyping members of a lower status group with concepts such as 'communal' and 'kind', plausibly because these traits are associated with the deferent behaviour higher status people expect from lower status people (Rudman and Glick, 2001). Thus, implicit biases involve semantic relations that are coherent and integrated because of the relation these semantic associations have with a social group's social-economic role (Jost and Banaji, 1994).

Nor does it appear legitimate to say that the relationship between different implicit biases is incoherent. Because the meaning of words that are associated with a social category match the meanings, explicit and implicit, that are associated with people with certain roles and status, different implicit biases are related to each other through tracking the relationships, and overlaps, between stereotyped roles. This allows us to explain why British people thought that immigrants to Britain were being described, when they were actually listening to a Slovenian stereotype of Bosnian immigrants. The low status of women and black people also explains why there is some overlap between the stereotypes of members of people from these social categories (Jost and Banaji, 1994).

Finally, we might think that while our implicit biases are integrated with each other, they are not integrated with our *stance* (Levy, 2011): Our stance being the set of values, beliefs and attitudes that form the roughly coherent standpoint from which we understand the world. But to assume that our implicit biases aren't part of our stance is an error, as we will show in the next section. Here we move on to the relationship between rationality and implicit biases, as we can understand the rationality of implicit biases precisely through an examination of the role they play in being an integrated part of our sense-making apparatus.

Rationality

Levy seems to understand the 'norms of reasoning' as the syntactical rules found in formal logic. Because such logic helps to prevent contradiction, applying it to our explicit beliefs causes them to be more coherent and consistent, he argues.

We do not seek to show that implicit biases are governed by syntactical operations. But we maintain that implicit biases are rational in three following ways:

1) They are sensitive to how the world is, and so change depending on input,
2) They are sensitive to moral reasons,
3) They are integral to how we make sense of the world.

While the first two types of rationality are arguably necessary for moral agency, for it would be hard to see how we could blame or praise someone for a mental process that wasn't sensitive to input, it is the third sense of rationality that we take to be most crucial for moral agency.

Before we explain the senses in which we think implicit biases are rational, it is worth pointing out that, while it may be true that explicit thinking is more rational than implicit biases in the sense of them being logical, formal logic is by no means endemic in our explicit reasoning. For example, it seems that we often base our explicit judgements on how familiar something is, what the majority believe, and what our friends believe (Haidt, 2001; Gigerenzer, 2008; Schwarz et al., 2007). It is not even clear that formal logic is always the most rational process, because such operations can be computationally intractable (Gigerenzer, 2008). However, we think it is fair to understand syntactical procedures as being unique to explicit reasoning – although we believe that this feature cannot contribute to the unified agent in the way Levy suggests, as we explain in part 2.

Importantly, reasoning according to syntactical rules is not the only way of understanding rationality that is relevant to moral responsibility. Rationality, in this sense, might e.g be understood as being responsive to reasons. Being rational in a way that allows us to be held morally accountable depends on an agent being responsive to *moral* reasons (Vargas, 2013). Vargas (2013) defines reasons responsive cognitive processes as responsive to (moral) reasons *iff* a person actually detects them in a particular situation, or would do in a suitable number of possible worlds, and those reasons either motivate a person to act in accordance to them, or would do in a suitable number of possible worlds.

On this account, an agent doesn't have to explicitly express something as a reason for it to be a reason: if a process is responsive to (moral) reasons, they are reasons for the

agent. When control for an action is not located in explicit reasoning processes, we can say only that conscious agency, but not necessarily a morally evaluable self, is bypassed, although Vargas (2013) suggests that conscious control may sometimes be necessary for moral agency.

Thus, Vargas (2013) argues that it is only those cognitive features that form part of the values and aims of the agent that are reason-responsive in the right way for which the agent can be held accountable. That is, for implicit biases to be the types of things that we can be held accountable for, they must be integral to how an agent qua agent makes sense of the world. We think implicit biases do play this role.

Before we discuss this further, it is worth noting that implicit biases are also reason responsive in two weaker senses. First, they change depending on how the world is, or appears to be. They are responsive to the roles people of particular social categories are given, and how they are presented to us through media. The roles and representations that are generally ascribed to members of a social category give us reasons to take such people as generally being 'a certain way'. For example, Dasgupta and Asgari (2004) found that implicit gender biases changed depending on how often college students had contact with women teaching counterstereotypical subjects, and after they had read counterstereotypical descriptions of famous women.

Secondly, insofar as the personhood and suffering of another is seen as a moral reason, it seems reasonable to suspect that implicit biases are responsive to moral reasons in this context as well. This is because implicit biases appear to be modulated by whether we are orientated to how another's life is for them. When white participants in an experiment took on the perspective of a black person, their implicit race biases decreased, their sensitivity towards injustices towards black people increased, and they were rated by black people as behaving more positively towards them (Todd et al., 2011).

In theory, perspective-taking works by allowing one to see members of a marginalised group as like oneself (Galinsky and Moskowitz, 2000). Furthermore, perspective taking can be triggered non-consciously (Davis et al., 1996), and therefore does not need to be explicitly guided. So we have reason to think that implicit biases are modulated when we attend to the experiences of others, something that is plausibly a moral reason.

Finally, and most crucially, implicit biases are rational insofar as they contribute to the coherent set of views and values an agent uses to make sense of the world. For example, consider the case of Joel, who has interviewed a variety of people for a job. To him, it is obvious that the best person for a job is candidate X. Candidate X is a 35 year-old white man and, to Joel, he came across as the most articulate, authoritative and competent. Furthermore, Joel didn't go with candidate Y because she wasn't as competent. So Joel can give good reasons for his choice. Candidate Y's gender had *nothing* to do with his decision. Joel knows this because he isn't sexist.

Yet, as can be shown with empirical evidence (Rooth, 2010; Rudman and Glick, 2001), it is likely that Joel's attitudes towards candidate Y were informed by his implicit biases. Candidate Y didn't appear as competent to Joel precisely because his implicit biases contributed to his tendency to take for granted that men are more

competent than women, and helped direct his attention and evaluations accordingly. His reading of the situation was organised by his tendency to value men more and to look for reasons why they deserve more credit.

This can be compared to other instances where we take peoples' actions to be indicative of implicit evaluations. For instance, Naomi forgets her daughter's football match because she was too busy with her work. Of course she cares much more about her child than her work, she can insist. Yet if it is a regular feature of Naomi's behaviour, we may begin to suspect that getting swept up in her work, and not attending to her daughter's needs, is indicative that Naomi implicitly evaluates her work as a higher priority than paying her child attention.

For socially sensitive beliefs, implicit biases have been shown to be better predictors of people's behaviour than their explicit beliefs (Greenwald et al., 2009b), and they are involved in judgements in a broad variety of contexts. Like with Naomi and her work values, implicit biases are involved in organising our attention and evaluation in numerous situations. Just as Naomi may be sensitive to the social norms that it is not acceptable for a mother to articulate that her work is, at least sometimes, more important to her than her child, and so she may not even believe herself how highly she values her work, or be able to articulate this, implicit biases can be seen as part of an agent's values and attitudes, even if they cannot articulate it.

Thus, while it is intuitive to some that only explicit reasons are reasons for the agent, it seems that some implicit processes play a similar role: they are receptive to reasons in that they change depending on input, including morally poignant input, and they are integrated with our stance as agents.

Implicit Stereotypes and Unification of Behaviour

Normativity is another supposed special feature of explicit attitudes that make us morally responsible for them. Levy (2012) argues that, "explicit attitudes integrate the agent in two ways: when they are employed in top-down reasoning, and when they activate the content they entail" (pp. 12). Now we will focus on the latter condition where 'normativity' means that "the dispositions associated with a belief are isomorphic with the dispositions we ought, normatively, to have" (ibid). For example, if we think 'I like chocolate cake', this causes us to seek out chocolate cake and to eat it when we find it. A coherent attitude will allow us to be unified through time if it causes us to act in ways that its content entails.

As already implied in the examples above, implicit biases appear to act as heuristics (i.e. rules of thumb) that normatively constrain our behaviour, consistently guiding how we evaluate others and how we act towards them.

For example, women who have implicit biases that associate maths strongly with men perform worse in maths tests (Kiefer and Sekaquaptewa, 2007). Their implicit biases normatively constrain their dispositions, in the sense that if one is aware of oneself as a woman, and is aware that women are not good at maths, then that gives one a reason not to be good at maths. That is, implicit biases have a normative effect here by contributing to the behaviour that their content entails. Further, Kiefer and Sekaquaptewa (2007) suggest that women with strong implicit gender-maths biases

have their implicit biases 'chronically accessible'. That is, this implicit bias will affect their behaviour in a consistent way whenever mathematic ability becomes salient to a situation.

It is plausible that many implicit stereotypes work in a similar way: some are implicated in guiding how you, as a member of a certain social category, *should* behave. In this way, then, implicit biases do appear to be involved in normatively constraining one's behaviour.

The content of implicit stereotypes also appears to influence our evaluations of how other people *should* behave, and our subsequent actions towards them[2]. Rudman and Glick (2001) have shown that implicit biases can predict how strong a backlash one will exhibit towards women characterised as 'dominant'. The backlash occurs because implicit biases dispose us to implicitly process women as people who 'ought' to be submissive, so we judge them at fault if they aren't. Furthermore, while one might understand the example of women and maths as a case where implicit biases are merely altering how we understand ourselves and the world to be, rather than a case of implicit biases acting normatively, this case is more clear. It is hard to explain why women are judged negatively if implicit stereotypes do not contribute to how we think they should behave.

Note that this interpretation of implicit biases also gives weight to the notion that they are involved in our sense-making capacities. If implicit biases normatively guide our behaviour then they act in a way similar to our explicit values, helping us to order the world, and guiding our attention, judgements and actions accordingly.

To summarise: Implicit biases systematically, and rationally, constrain and influence our dispositions. We have shown that they are coherent and integrated and that they rationally constrain our behaviour in a coherent and integrated way.

Implicit biases, it turns out, do not lack structure, can be part of our stance, and do normatively constrain our dispositions. They contribute to an agent's behaviour, judgements, and stance in a way that does not undermine their consistency through time.

PART 2 Assertibility and unified agents

Why assertibility matters

We started this paper with the observation that it seems plausible to accept that explicit attitudes are special when it comes to moral responsibility. We then discussed one reason to think that our relationship to our explicit attitudes is different to the one we have with our implicit ones – the notion that explicit attitudes, in contrast to implicit ones, unify us as agents. But by the end of our investigation, not much has

[2] While this interpretation of the literature may be up for debate, there is much literature on how stereotypes perform a prescriptive function (for example, see Fiske and Stevens (1993) and Burgess and Borgida (1999)).

survived of this idea. Implicit stereotypes seem just as central in constituting us as agents as do explicit ones, when it comes to broad content, rationality and normativity.

Nevertheless, we have not yet discussed what one might think is the most important reason for the intuition that there is something special about our explicit mind – namely, the idea that we do have a special kind of control over our explicit mind that we do not have over our implicit attitudes – a belief that the former but not the latter is available for the effortful syntactical manipulation of propositional thought. We have so far not engaged with this argument, which we take to be at the heart of Levy's argument. The remainder of the paper addresses this issue and looks at some reasons why this assumption, although intuitive, might not be correct.

In order to do this, first of all, we need to make it clear exactly what we do have in mind, because not all attitudes that we are conscious of are ones that can be controlled in this specific way. You might well be aware of your spider phobia, but this does mean that simply deliberating about the fact that there are no dangerous spiders in the Northern hemisphere will make your phobia go away. Thus, it is not just the fact that you are aware of having an attitude that makes it a candidate for the specific control that we are interested in here.

Levy distinguishes between attitudes like phobias and ones that are controllable in the right kind of way. Levy says that only these kinds of attitudes, but not phobias and not implicit attitudes, can be expressed in propositional form and can form the premises in a rational argument.[3] This difference matters according to Levy, because it is these two characteristics of explicit attitudes that explain why these attitudes have a special role in integrating the agent. As we have already discussed, an integrated agent seems to be a key *desideratum* for any morally responsible agent.

So what does Levy mean when he says that some explicit attitudes can be expressed in propositional form and can form the premise in a rational argument? The example of the spider phobia will help to clarify this. In the case of such a phobia, the sufferer will often be very aware of the fact that they are suffering from it and often they will know also how they are disposed to react in the presence of a spider. They will even know what kind of mental states the presence of a spider will produce in them. They know that they will feel fear and experience the spider as threatening and dangerous.

Nonetheless, even though they possess all that knowledge, they would not be able to sincerely assert that they believe that spiders are dangerous. Even though the spider produces such a comprehensive set of integrated emotional states and response dispositions, it does not produce the equivalent assertible belief. On the contrary, the typical assertion from a phobia sufferer will be that they know that spiders are not

[3] Levy presents this distinction as a response to King & Carruthers (2012), who argue that cognitive science shows that we are never directly conscious of our attitudes and that they should therefore not have a special status for moral responsibility. Levy's response is to assert that this special status is not attributable to the direct access but to the functional role attitudes play when we are conscious. We agree with that argument in principle, but we are not sure that Levy has demonstrated that explicit states really do play the functional role that he attributes to them.

dangerous in the northern hemisphere, but that they can't help feeling threatened by them anyway.

According to Levy, it is exactly this ability to sincerely assert the content of a mental state that makes some explicit states special for moral responsibility, and it is exactly the lack of this assertibility in the case of phobias that explains why they are not special in this way, even if the subject is conscious of having them (Levy 2012, page 12).

So the question now has become, why does assertibility matter? It matters according to Levy, because only asserted beliefs can be, as it were, cast in a different medium (language), and this new medium allows a whole set of operations that are only possible in language.[4] Obviously, the agent can also put the content of the phobia in a sentence in natural language, but the important difference is that in this case the agent does not thereby express the phobia[5], and because of that, will not feel that they have to believe anything either that follows rationally from the proposition contained in the sentence. This is in sharp contrast to the case where the agent can assert their belief sincerely. In this case, as the assertion expresses their belief, they are bound to accept all the logical entailments that can be made explicit by manipulating the assertion according to the laws of logic. This in turn has a hugely unifying effect on the beliefs an agent will hold, because the fact that they are manipulable means that it is far easier to bring individual beliefs in contact with each other and to spot potential incoherencies.

What is the problem?

The first problem with this account is that it invites us to buy into a false dichotomy. It seems very plausible to say that attitudes like a spider phobia are not part of the responsible agent in the same way as let's say an explicitly endorsed fear of criminals. The phobia might well have a powerful influence on the behaviour of the agent, but it does not play a role in how the agent understands the world, insofar as the agent does not feel that the assertion of the attitude expresses her world view. She therefore would not feel the need to give reasons for why it is right to have this attitude, because rightness is not a criterion for holding it. Compare that to the explicitly endorsed fear of criminals. Here the agent claims that she is afraid of criminals because she has reasons for that fear. In asserting her fear, she claims that there are criminals out there and that she has reasons to fear them.

So the phobia case is intuitive, because here the gut feeling is not only clearly in contrast to what the agent would want to assert, it also is extremely insensitive to evidence. But many of our gut feelings are not like that on either of these two dimensions. As we already discussed in the first part, implicit stereotypes are surprisingly sensitive to different inputs (see the section on rationality in part one). So

[4] The importance of language as a tool to make content visible and manipulable is a very influential one in philosophy of mind. See e.g. Dennett (1998), Clark (2006), & Pettit & McGeer (2002), reference suppressed.

[5] i.e. they would not agree that the proposition 'spiders are dangerous' contained in the sentence is true.

the spider phobia is not like implicit stereotypes in this respect, because they do co-vary systematically with different environments, whereas a phobia does not. But is it not at least true that gut feelings are different to explicitly assertible beliefs, because they do not express our beliefs?

This seems very plausible, but on a closer look it turns out to be far less obvious than we might have thought. The phobia case is quite special, because it is one of the few cases where the agent is very sure that the gut feeling experienced is just about their psychology, but not about how the world really is. In most cases where gut feelings compete with judgments that we have formed on the basis of explicitly available reasons, that is not the situation we find ourselves in. Famously[6], Huck Finn's conscious deliberation leads him to the judgement that he should betray the runaway slave Jim to his owners, but at the same time Huck has the strong gut instinct that there is something wrong with this course of action. Huck does not have reasons for this feeling that he can express, but this does not mean that he views his gut instinct like a phobia. Instead, he in the end decides to act in accordance with his gut feeling, even though he is fully aware that by doing so he is violating what follows from the explicit reasons available to him, because he feels that what his reasons tell him to do is in some indescribable sense wrong. [7]

Huck's balance of assertible reasons for or against turning Jim in leads him to the judgement that he ought to return Jim to his rightful owner. But this does not mean that he can assert that he is convinced that it is overall the best thing to act according to the conclusion he can reach by deliberation on his assertible reasons. The strong gut feeling that something is wrong with his reasons leads him to decide to act against the conclusion derived from those reasons and in line with his gut feeling. Thus, his gut feeling is in an important sense an expression of a deeply held conviction, even though it does not at all cohere with the assertible reasons available to him.

Another more mundane example is the fireman Richard Holton discusses in his essay on choice (2008). While working in a burning house, this fireman suddenly has the strong sense that he urgently ought to leave. He works through all the reasons that he has available from his training to find out what might be wrong, but cannot find anything. However the feeling gets stronger and stronger and the fireman decides to trust his feeling and leaves the house. In the story this saves his life, because the house collapses seconds after he leaves it. What might have happened here, and what might happen in similar situations, is that he unconsciously picked up on a feature of the situation, but was not able to express it in an explicit assertion.

This example neatly shows that it is not only literary cases like Huck Finn where agents find themselves torn between judgements reached by evaluating assertible

---

[6] See e.g. Arpaly (2002) for an in depth discussion of the philosophical relevance of the Huck Finn case.

[7] This leads Richard Holton (2008) to use Huck Finn as an example for a distinction between weakness of the will and akrasia that he favours. According to this distinction, Huck Finn is weak willed, because he does not follow through on his resolution to turn Jim in, but he is not akratic, because by the time he has to execute his resolution, he is not convinced any more that it is the best thing to do, even though he still does not have good assertible reasons why it is not.

reasons and a gut feeling. In these situations, it is very often not clear to the agent whether her gut feelings describe the world accurately.

However, even though it seems to be true that the clear distinction between assertible judgments and mere describable gut feelings is less clear than Levy seems to think, this fact on its own only does limited damage to Levy's claim. It still seems to be true that sincerely asserted beliefs create strong normative pressure to accept their entailments, and this still seems a unique feature of explicit beliefs and a powerful tool to generate unified agents. Gut feelings might well express something important about the agent, but they don't seem to lend themselves to deliberation. As the reasons that the agent has for holding them are not explicit, she cannot integrate them easily into her web of beliefs.

One way to counter this argument is to point out that this fact only seems to matter if it really is the case that agents normally act on the conclusions they reached after conscious deliberation. But if it were the case that conscious deliberation does not normally play an important role in action guidance, then Levy's point looks more vulnerable.

We have already provided many reasons in the first part of this paper for why it might be the case that we hugely underestimate the importance of our implicit states in behaviour control. At the same time, another reason to think that the role of explicit attitudes in the control of behaviour might be smaller than commonly assumed comes from the literature on confabulation (e.g. Wegner 2002, Wilson 2002). This literature demonstrates that in very many cases people are not aware of their real motives for acting and use explicit reasoning only post hoc to rationalise their behaviour. So perhaps the impression that explicit reasons are action-guiding comes from such rationalisations. In that case, it would not be surprising that we accept belief and entailments sincerely, because we only rationalise what our explicit belief is that entailed the action after we performed the action for unconscious and often entirely unrelated motives.

Whether this really is how we should think of the role of conscious deliberation in action guidance is obviously disputed (it is worth noting here though, that this literature was the reason for Carruthers (Carruthers, 2007) to doubt the existence of conscious action guidance). At this point, all that we can assert with confidence is that even though it is true that explicit content can be manipulated in a way that is not available for implicit content, and even though it is also true that manipulating expressed beliefs in this way does create normative pressure to accept the conclusion of the deliberation, this does not mean that a) gut feelings do not express our understanding of the world, nor does it mean b) that conscious deliberation does play a crucial role in action guidance.

Control by effortful reasoning

However, Levy might not be too worried about this last point. Whether or not conscious deliberation really leads to action, it does at least lead directly to the acquisition of a new belief. It might well be that there are psychological reasons why this so-acquired belief does not always control actions, but it nevertheless is a belief that will help to integrate the overall agent. That deliberation really does create new

beliefs might seem obvious. After discussing Levy's account of assertions, it seems that we now also have a philosophical explanation for this intuition. The control we have over our beliefs arrived at by deliberation stems from the fact that deliberation is about performing effortful syntactical operations on our explicit beliefs asserted in language. This effortful control exercised in deliberation seems nothing else than intentional action.[8] As Levy writes: Only the latter [effortful reasoning] is an exercise of agency: something we do, rather than something that happens to us.

But even though this seems very intuitive, once we look closer it is not so clear what this agential control actually is. The act of deliberation, as we have seen in the discussion of our gut feeling cases, contains two quite distinct episodes. The first part is the execution of the logical operation (in our case, the entailment from all slaves should be returned to their owners to this slave should be returned to his owner). So Huck could work out by applying the laws of logic that he should turn Jim in, on the basis of the beliefs that he had expressed. Nevertheless, as we saw, that did not mean that he was prepared to accept the conclusion of this operation. Interestingly though, this non-acceptance of the conclusion was not something where Huck intentionally applied any logical rules, or in fact performed any intentional actions at all. Huck evaluated the proposition negatively, but this evaluation, if it is an act at all, does not seem to be an intentional action. In other words, it does not seem to be the case that the agent can decide at will how to evaluate. The result of the evaluation is driven by what the agent takes to be the case, rather than what he wants.[9]

If this distinction holds, then the acquiring of his new belief takes place in two stages. First, there is the act of applying the rules of the calculus in inner speech and then there is a second step, where the result of this operation is either accepted or not. In Huck's case it is not, but the important point here is that these two steps are not only characteristic of Huck's decision making, but of all the decision making there is. Very often we do not notice the second step, because after going through the calculus it seems obvious that the results have to be accepted. In fact these two different steps are very much what motivated Carruthers (2007) originally to claim that we are not directly aware of our attitudes. We do directly know the result of our deliberations, but we do not know directly whether we have convinced ourselves.

The point is that we need to separate the intentional setting up and applying of the calculus[10] from the evaluative act of accepting the result as the solution to our question.[11] Once we accept that this is a distinction that separates two distinct mental

---

[8] The terminology of "effortful reasoning" is strongly reminiscent of Baumeister and indicates that Levy here, like Baumeister (e.g. 2007), works in a two-systems architecture and identifies the special process of explicit effortful reasoning as a system 2 process.

[9] We cannot discuss this in any depth here, but there is a large literature on mental actions which emphasises this point. See e.g. Hieronymi (2009), Mele (2009), Strawson (2002), reference suppressed.

[10] Actually, there might be two evaluative acts here. You can set up the calculus intentionally, but perhaps actually applying the rules is already evaluative. It is then a further evaluative act to accept the outcome not only as a correct application, but also as factually correct.

[11] See e.g. Hieronymi (2009), Mele (2009), Strawson (2002), reference suppressed.

episodes from each other, we can now ask ourselves whether the individual episodes are suitable candidates for the direct control that according to Levy makes explicit beliefs special.

Let's have a look at the first episode first. This is the intentional action of setting up and applying the calculus. For these episodes, it is clearly true that we have full intentional control over them. We decide which operation we want to use and then intentionally try to fit the content under discussion in the schema we want to use. So the control element is there, but as we have seen, what we control here is simply the syntactical manipulation of bits of inner speech. In order to actually create a new attitude, we need the evaluative act. However, the problem with the evaluative act separated out from the intentional part of the deliberation is that it works quite similarly to an automatic process like a gut feeling. What the agent judges to be correct is not under her intentional control.

But if this is right, then Levy now has a problem. Levy claims that asserted content is special, because the agent has intentional control over it, which leads to a unified agent because it connects beliefs in a way that implicit stuff could not. But if we accept that it is not the effortful manipulation of beliefs that directly leads to new beliefs, then it looks now as if the actual production of a new belief that is coherent with the agent's belief system is not something that the agent has intentional control over either.

But if Levy opts for the first part of the process and claims that what is special about explicit content is simply the fact that it can be manipulated, then it becomes difficult to see what the difference between propositions that merely describe beliefs and propositions that express beliefs is supposed to be – as both of them can be manipulated in exactly the same way.

Thus, in the first case, it really is true that the evaluative step does produce new attitudes that are likely to be more coherent with the belief system, but it is unclear whether the process that leads to that result really is controlled in a different way from implicit processes. In the second case, on the other hand, it is true that the explicit process is clearly controlled in a different way to the implicit ones, but here the results of the manipulation do not lead to any change in attitude on their own.

Conclusions

Explicit content is special because language makes it intentionally manipulable. It is not so clear, however, that the same is true for the attitudes themselves.

Whether we acquire a new attitude or not does not seem to be a question of direct intentional control, but of an evaluative act. This at least is the message from a large literature that argues that looking for intentional control is looking for the wrong kind of control when trying to understand belief acquisition. Instead, what does matter is that the attitudes are reason responsive, or judgment sensitive (Vargas 2013, Holroyd, forthcoming) – but judgment sensitivity is something that applies not only to explicit attitudes, but also to very many implicit ones, as the discussion in the first part of the paper demonstrated.

We do not want to claim that we have shown that there could not be a principled difference between conscious evaluative acts and implicit ones. We do however claim that Levy has not shown us what the difference might be, because the type of intentional control that is specific to explicit content he discusses is not evaluative control at all.

On the other hand, we do not want to rule out Levy being right in claiming that there is a special connection between the practice of responsibility ascription and the intentional manipulation of explicit content. In fact, one of us has said a lot about this connection elsewhere (reference suppressed). Here, all we have shown is that Levy's claim about the intentional manipulation of expressed beliefs does face severe problems, because it is not clear that he gets the direct control from this manipulation, which seems to be his aim.

Explicit intentional manipulation of content in inner speech, without evaluative acts, so we argue, might not reveal much about our moral stance as agents. Even if we accept Levy's point that there is a clear difference between explicit content that is sincerely asserted and explicit content that is merely described, as in the case of the phobia[12], this does not entail in any way that agents will be convinced by the results of conscious syntactical manipulations of sincerely asserted content.

There is no one privileged way to work out what this stance is, and it is well possible that our hunches and our behaviour say more about us as a moral agent than our explicitly made sincere assertions.

Literature:
(references to own work suppressed)

Arpaly, N. (2002). *Unprincipled virtue: An inquiry into moral agency*. Oxford University Press.
Avenanti, A., Sirigu, A., & Aglioti, S. M. (2010). Racial bias reduces empathic sensorimotor resonance with other-race pain. *Current Biology*, *20*(11), 1018-1022.
Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions In Psychological Science*, 16(6), 351-355
Burgess, D., & Borgida, E. (1999). Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, Public Policy, and Law*, *5*(3), 665.
Carruthers, P. (2007) The illusion of conscious will. *Synthese* 159: 197-213
Clark, A. (2006). Material symbols. *Philosophical Psychology*, 19(3), 1–17.
Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*(5), 642-658.
Davies, M. and Stone, T. (Eds.) (1995). *Folk psychology*. Oxford: Blackwell.

[12] And the paper has provided reasons to be at least very careful here.

Davis, M. H., Conklin, L., Smith, A., & Luce, C. (1996). Effect of perspective taking on the cognitive representation of persons: a merging of self and other. *Journal of personality and social psychology*, *70*(4), 713.
Dennett, D. (1998). Reflections on language and mind. In P. Carruthers, and J. Boucher (Eds.) *Language and thought: interdisciplinary themes*, pp. 284–94. Cambridge: Cambridge University Press.
Dovidio, J. F., Gaertner, S. E., Kawakami, K., & Hodson, G. (2002). Why can't we just get along? Interpersonal biases and interracial distrust. *Cultural Diversity and Ethnic Minority Psychology*, *8*(2), 88.
Fiske, S. T., & Stevens, L. E. (1993). *What's so special about sex? Gender stereotyping and discrimination*. Sage Publications, Inc.
Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of personality and social psychology*, *78*(4), 708.
Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford University Press.
Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, *94*(4), 945-967.
Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009a). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of personality and social psychology*, *97*(1), 17.
Greenwald, A. G., Smith, C. T., Sriram, N., Bar- Anan, Y., & Nosek, B. A. (2009b). Implicit race attitudes predicted vote in the 2008 US presidential election. *Analyses of Social Issues and Public Policy*, *9*(1), 241-253.
Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.
Hieronymi, P. (2009). Two kinds of mental agency. In L. O'Brien and M. Soteriou (Eds.) *Mental Actions*, pp. 138–62. Oxford: Oxford University Press.
Holton, R. (2009). *Willing, Wanting, Waiting*. Oxford: Oxford University Press.
Hoffman, C., & Hurst, N. (1990). Gender stereotypes: Perception or rationalization?. *Journal of Personality and Social Psychology*, *58*(2), 197
Hutto, D. (2008) *Folk Psychological Narratives. The sociocultural basis of understanding reasons.* Cambridge MA: MIT Press.
Jost, J. T., & Banaji, M. R. (1994). The role of stereotyping in system justification and the production of false consciousness. *British Journal of Social Psychology*, *33*(1), 1-27.
Kiefer, A. K., & Sekaquaptewa, D. (2007). Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility to stereotype threat. *Journal of Experimental Social Psychology*, *43*(5), 825-832.
King, M., & Carruthers, P. (2012). Moral responsibility and consciousness. *Journal of Moral Philosophy*, *9*(2), 200-228.
Levy, N. (2011). Expressing who we are: moral responsibility and awareness of our reasons for action. *Analytic Philosophy*, *52*(4), 243-261.
Levy, N. (2012). Consciousness, implicit attitudes and moral responsibility. *Noûs*.
Mele, A. (2009) Mental actions a case study. In O'Brien, L. & Soteriou, M. (Eds.) *Mental Actions*, pp. 17–37. Oxford: Oxford University Press.
Moran, R. (2001). *Authority and Estrangement: An Essay on Self-knowledge.* Princeton, NJ: Princeton University Press.

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101.
Pettit, P. and McGeer, V. (2002). The self-regulating mind. *Language and Communication*, 22, 281–99.
Project Implicit. (2011.) Web. Last accessed: 2 July 2013. <https://implicit.harvard.edu/implicit/>.
Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523-534.
Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of social issues*, *57*(4), 743-762.
Schmid Mast, M. (2004). Men are hierarchical, women are egalitarian: an implicit gender stereotype. *Swiss Journal of Psychology*, *63*(2), 107.
Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in experimental social psychology*, *39*, 127-161.
Stanley, D. A., Sokol-Hessner, P., Banaji, M. R., & Phelps, E. A. (2011). Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proceedings of the National Academy of Sciences*, *108*(19), 7710-7715.
Strawson, G. (2003). 'Mental Ballistics Or The Involuntariness of Spontaneity.' *Meeting of the Aristotelian Society*, University of London, 28 April.
Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review*, 117, 245–73.
Todd, A. R., Bodenhausen, G. V., Richeson, J. A., & Galinsky, A. D. (2011). Perspective taking combats automatic expressions of racial bias. *Journal of personality and social psychology*, *100*(6), 1027.
Vargas, M. (2013). Situationism and moral responsibility: free will in fragments. *Decomposing the Will*, 325-49.
Wegner, D. M. (2002). *The Illusion of Conscious Will*. Cambridge MA, MIT Press.
Wilson, T. (2002). *Strangers to Ourselves; Discovering the Adaptive Unconscious*. Cambridge MA: Belknap Press.
Zawidzki, T. (2008). The function of folk psychology: mind reading or mind shaping? *Philosophical Explorations*, 11(3), 193–210.